Modeling Those F-Conditions – Or Not

Richard Schwartz, Hubert Jin, Francis Kubala, †Spyros Matsoukas

BBN Systems and Technologies, Cambridge MA 02138
†Northeastern University, Boston MA 02115

ABSTRACT

After several disappointing preliminary attempts to make condition-specific models, we decided that it would be more advantageous to spend our time just trying to improve the core recognition system and use general adaptation techniques to deal with variations. This simplified the system immensely and freed up people and also made the transition from the PE system to the UE system much easier. We describe our attempts at condition-specific modeling/adaptation/training. We show that the benefit for channel-specific modeling is smaller than that for general adaption procedures and we argue that the cost is too high.

1. INTRODUCTION

In the November 1993 Wall Street Journal (WSJ) [1] evaluation we learned that if we wanted to use wide-bandwidth training speech to recognize telephone speech, it was essential to bandlimit the training speech. This was more effective than the various adaptation methods (although adaptation provided a small additional gain). During the 1995 MarketPlace evaluation, IBM [2] showed a significant gain for segmenting and classifying the input into three categories: clean, telephone, and speech with music. For each condition, they used four steps:

1. preprocess all of the WSJ training data to be more like the condition,
2. use supervised adaptation with the specific data provided,
3. during recognition classify the input into the appropriate category,
4. use the appropriate model and then use unsupervised adaptation.

For example, the training was bandlimited for the telephone condition, while for the speech plus music condition, music was added to all of the WSJ training speech. This approach was based on the assumption that there was not enough training data of each type, so the WSJ data must be used.

There were several differences in the evaluation this year. The speech came from many different shows, there was more training speech available, and many more conditions with subjective labels were identified. In particular, there were five binary attributes: spontaneous speech, degraded channel, music, noise, nonnative accent. Many of the 32 possible combinations actually occurred in the data. An attempt was made to define several focus (F) conditions, but each was necessarily an aggreagation of a few of the 32 possible combinations.

It is tempting to develop specific solutions for each of these F-conditions, or preferably for arbitrary combinations of the binary features. However, we preferred not to do this for several reasons:

1. This can result in a complicated system with many models and multiple sequential decisions.
2. It requires separate research effort for each condition as well as for how to detect and combine them.
3. Most importantly, while some small gains were possible for some conditions, we found that the overall gains were small.

1.1. Alternative Approach

Instead, this year [3] we adopted an alternative approach in which we

1. Use general technology for all conditions. This frees up people to work on the general problem.
2. Use the available Broadcast News training data, without resorting to the modification and use of other (e.g., WSJ) data.
3. Use adaptive training to remove the differences peculiar to each condition (without having to label them).
4. Use clustering and unsupervised adaptation during recognition to model any combination of effects as well as any new effects.

In Section 2 we describe our attempts to deal with some specific conditions. In Section 3 we present results showing the gains for each condition obtained by using general adaptive training and recognition.

2. CONDITION-SPECIFIC MODELS

In this section we discuss some experiments aimed at modeling particular features of different kinds of speech. In particular, we attempted to preprocess the training data to model telephone speech, to train the system on an appropriate subset of the data for each condition, and to adapt the model trained on all of the speech to each labeled F-condition.

The tests described here were all performed under the following conditions in order to simplify the experiments and also in order to obtain an upper bound on the improvements that could be expected from condition-specific models.

First, we trained and tested in PE (Partitioned Evaluation) mode. Thus, even on the test, we knew the condition of the data, and also new the start and end of individual speakers turns. This provided an upperbound on the gains for using condition-specific models, since in the real UE (Unpartitioned Evaluation) condition, we would expect to make segmentation and classification errors on the conditions.

Since most of the training speech was not available until just before the actual evaluation, we used a set derived from the first release of data (nominal 30 hours), which contained about 16 hours of usable training speech. Again, this provided an upper bound on the improvements, since we might expect that if we had sufficient appropriate training speech, data conditioning and normalization techniques would be expected to help less.

Third, we tested using a simplified system based on phonetically-tied mixture (PTM) densities. We also tested under known gender conditions to further simplify the experiments.

2.1. Bandlimited Telephone Models

We made several attempts to model telephone speech. The first attempt was based on simply bandlimiting all the 16 hours of training speech. We bandlimited the training data during the spectral analysis (as we had done in the past). The analysis in BYBLOS [9] first computes a power spectrum of the signal. Then we resample the power spectrum nonuniformly to apply a Mel-scale weighting. During the resampling, we restrict the frequencies used to those desired (normally 80 Hz to 6 kHz). Then we apply an inverse cosine transform to the resampled spectrum to produce cepstra. This mechanism allows us to further restrict the bandwidth to any desired range.

In our experiments, the same analysis was performed on training and test data. (This is one of the complications of using condition-specific models. It means that if we identify some test speech as being telephone-like, we must then redo the spectral analysis from scratch.)

Table 1 shows the error rate for different band limits for the F0 and F1 speech (both clean and wideband), for F2 (low fidelity), and for all conditions. One problem with the F-condition labels is that much of the speech labeled as F2 is clearly not telephone speech. It is degraded in some other way. Thus, it is not clear what gain can be expected. Rather than implement a bandwidth detector, we simply measured our performance on one particularly difficult episode in which F2 appeared to consist primarily of narrowband speech (based on long term spectral plots). This is indicated as the "telephone" condition in the table.

From this table we see that the usual band levels associated with telephone audio (300-3400 Hz) degrade all conditions. When we used a slightly wider band (125-3750 Hz) there was still no gain on the F2 condition, but there was a small gain on speech that is clearly from the telephone. We also considered using wider bandlimits (80-7500 Hz). We can see that there is a very small gain on clean speech and a modest loss on the F2 condition.

	Frequency Range			
From	80	300	125	80
Condition To	6000	3400	3750	7500
F0. prepared	19.1	27.1	22.0	19.0
F1. spontaneous	42.9	50.5	46.1	42.6
F2. low fidelity	51.2	52.5	51.2	51.8
true telephone	63.2		61.1	
OVERALL	39.8	47.5	43.0	39.6

Table 1: Error rate with different bandlimiting on training and test.

Thus, we no longer observe a large advantage on telephone test data for bandlimiting the training data. It is worth commenting here that the actual telephone condition accounts for a small portion (probably less than 10%) of the test data. Even if we had managed to cut the error on this data in half relative to just using one model it would only decrease the overall error rate by less than 5% at a significant cost – in terms of system complexity.

2.2. Condition-Specific Training

We know that it is possible to reduce recognition error somewhat by making separate models for male and female speech. That is, the improvement for making more specific models is larger than the loss due to discarding half of the training speech. Depending on the system, researchers report anywhere from 5% to 15% reduction in error rate relative to a gender-independent model. (This gain may disappear if speaker adaptation is used during the training.)

The question is whether the F-conditions are sufficiently different that fragmenting the training data will be advantageous. One measure of how different conditions are is the increased error rate when we train on one condition only and test on another. For example, we know that a model trained on one gender and tested on the other has several times the error (for low error rates) as one trained on the correct gender. The difference in genders is large enough that adding an equal amount of the wrong gender to the model training – without adaptation – results in a slight degradation. However, when testing on one of the degraded conditions, the cost for training only on clean read speech, relative to training only on the degraded condition is less than a factor of two. An alternative question is when testing on clean read speech, if you already have some similar training, will adding in the various forms of degraded speech help or hurt?

We made a condition-specific model for "clean" speech from all of the data marked as F0 (clean, wideband, read, native) and F1 (+spontaneous).

In Table 2 we measure the effect of training the model with only F0 and F1 data (i.e., discarding the other 50% of the data). In both conditions, we did not use adaptation on the test data. Clearly adaptation on the test data would help more in the case of condition-independent models.

	Training Data	
Condition	All	F0,F1
F0. prepared	19.1	20.0
F1. spontaneous	42.8	42.7

Table 2: Error rate on clean wideband speech (F0 and F1) when training on all speech vs. only on F0 and F1 speech.

The results show that it is better to include the data from other conditions than it is to discard it.

2.3. Supervised Condition Adaptation

There are many different conditions for which we may not have specific algorithms. A more general approach would be to adapt a model trained on all of the speech to each of the marked conditions using supervised adaptation. This only requires training the system once, and quickly produces many condition-specific models. Since the training is supervised and there is substantial training for each condition, the adaptation can be quite detailed.

We show the improvement obtained with supervised condition adaptation in Table 3.

Adapt Training Adapt Test	NO NO	YES NO
Condition	SI	Sup
F0. prepared	16.6	16.1
F1. spontaneous	39.4	37.8
F2. low fidelity	45.4	44.5
F3. music	32.0	30.8
F4. noise	25.6	24.8
F5. non-native	30.8	31.0
FX. mixed	58.4	57.4
OVERALL	35.2	34.3

Table 3: Condition-Adapted models. The SI model was transformed by supervised adaptation to each F-condition in the test.

We see a small but fairly consistent reduction in error across all conditions. On the average the error rate is reduced by 0.9% absolute or about 3% relative. Of course, we could not expect this full advantage, since this method also requires that we be able to determine the appropriate model to use during recognition.

2.4. Using Condition-Specific Models

If we had been successful at making large reductions in error rate by using condition-specific models, then we would also be obligated to develop algorithms for segmenting and classifying a passage into the appropriate condition. (As we said earlier, we would ideally determine which of the binary attributes were present, and then construct a model for the combinations.) Leaving aside the more difficult segmentation problem, we attempted to build a simple classifier for the F-condition using a "speaker-identification" type algorithm [4]. Our initial attempts at seven-way classification were not satisfactory.

An alternative to preclassifying the segments would be to recognize each passage with multiple models. However, this alternative is clearly unnattractive, because it is computationally expensive, and it is not clear how this would work on data that was not already segmented. Also, it does not allow for multiple simultaneous independent condition features.

3. CONDITION-INDEPENDENT MODELS

There are clearly several advantages for using a single model and procedure for all conditions.

1. It is not necessary to segment or classify the passage.
2. It is only necessary to estimate one model.
3. We can concentrate all of our research effort on methods that improve all conditions.

One of the general tools we have at our disposal is adaptation (really normalization) techniques. There are various parameters and methods for adaptation. The most commonly used approach is to adapt the model, unsupervised, to the test speech using multiple passes of recognition. We used the MLLR techniques of Legetter, et. al. [5]. We found that if we used more than two transformations, the system memorized the recognition errors of the first pass. We improved the unsupervised adaptation on the test by clustering together different segments that appeared to be from the same speaker on the same channel [6]. We also considered supervised adaptation of the model to a known channel condition as described in the previous section. In this case, we used more transformations because there was more data and the transcriptions were known. Finally, we used SAT [7]. This can be thought of as removing the characteristics of each speaker from the training. The technique we used [7] actually finds the "compact" model that results in the highest likelihood for all the speakers' data, given their corresponding transformations. We have also developed a more practical method that literally just removes the differences before combining speakers in the training [8].

We show below in Table 4 the results obtained for each condition with various combinations of adapting the model to the condition, unsupervised speaker-adaptation on the test, and Speaker Adaptive Training (SAT).

Adapt Training Adapt on Test	NO NO	YES NO	NO YES	YES YES	SAT YES
models/gender	1	7	1	7	1
Condition	SI	Sup	SA	SupSA	SATSA
F0. prepared	16.6	16.1	15.3	14.9	14.8
F1. spontaneous	39.4	37.8	36.7	35.2	35.1
F2. low fidelity	45.4	44.5	40.1	40.4	40.2
F3. music	32.0	30.8	30.2	29.6	30.2
F4. noise	25.6	24.8	24.1	23.4	25.0
F5. non-native	30.8	31.0	25.9	25.6	23.4
FX. mixed	58.4	57.4	54.8	54.0	53.7
OVERALL	35.2	34.3	32.3	31.7	31.6

Table 4: Word error rate by test condition, for SI, supervised condition adaption, unsupervised adaptation on test, supervised condition adaptation plus unsupervised adaptation on test, and SAT adapted training with adaptation on recognition.

We can see that the gain from unsupervised adaptation to the test (3rd column of results) is significant and uniformly larger than that from adapting the model to the condition (2nd column). The results in the 4th column show that there is a small (average 0.6%) gain for adapting the model to the known condition before unsupervised adaptation to the test. Finally, the last column shows the results for SAT. While this gain is also not large, it comes with no added cost during recognition. There is only one model (per sex) and no need to determine the condition before recognition. It is also interesting to note that the F2 condition that we tried so hard to fix is improved significantly simply by using general adaptation techniques.

4. DISCUSSION

We do not dispute that further progress could be made on condition-specific modeling. Rather we believe that, even if we had achieved a 10% overall gain by these methods, the costs (in complexity and computation during recogniiton, and fragmented research areas) is too high. In addition, the results are limited to the performance of the core system on clean careful speech, which is still well over 10% word error.

For the coming year we plan to focus only on methods that apply to all of the conditions. In particular, our experience (and the results across sites in this evaluation) show that improvements to the clean speech condition carry over to all conditions.

We used gender-dependent models this year, but we will switch to a single model shortly in order to remove the last requirement for condition detection.

5. CONCLUSIONS

We made several attempts at developing condition-specific models, but the gains were quite small. We found that the gains from general adaptation techniques, applied both during training and recognition were significantly larger, and resulted in a much simpler system. We did not rely on any training speech from other corpora. The resulting system was relatively easy to run and resulted in quite good performance during the formal evaluations.

Acknowledgements

This work was supported by the Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. Anastasakos, T., F. Kubala, J. Makhoul, R. Schwartz, "Adaptation to New Microphones Using Tied-Mixture Normalization", *Proceedings of ICASSP-94*, Adelaide, South Australia, Apr. 1994, vol. 1, pp. 433-436.
2. Gopalakrishnan, P., R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, H. Printz, M. Franz, "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System", *1997 DARPA Speech Recognition Workshop*, Harriman, NY, Feb. 1996, pp. 72-76.
3. Kubala, F., H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, J. Makhoul, "The 1996 BBN BYBLOS Hub-4 Transcription System, *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
4. Gish, H., M. Siu, R. Rolicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proceedings of ICASSP-91*, Toronto, Canada, May 1991, vol. 1, pp. 701-704.
5. Leggetter, C. J., P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of the Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 110-115.
6. Jin, H., F. Kubala, R. Schwartz, "Automatic Speaker Clustering", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
7. Anastasakos, T., J. McDonough, R. Schwartz, "A Compact Model for Speaker-Adaptive Training", *Proceedings of ICSLP-96*, Philadelphia PA, Oct. 1996.
8. Matsoukas, S., R. Schwartz, H. Jin, L. Nguyen, "Practical Implementations of Speaker-Adaptive Training", *1997 DARPA Speech Recognition Workshop*, Chantilly VA, Feb. 1997, elsewhere this volume.
9. Nguyen, L., T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos, Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System", *Proceedings of the Spoken Language Systems Technology Workshop*, Austin TX, Jan. 1995, pp. 77-81.